基于图聚类与蚁群算法的社交网络聚类算法 *

叶小莺[1]，万 梅[2]，唐 蓉[3]，谢 云[1]，陈桂宏[4]，李 强[1]

(1. 广东东软学院 计算机科学与技术系，广东 佛山 528225; 2. 广州工商学院 计算机科学与工程系，广州 510850; 3. 重庆市九龙坡区精神卫生中心，重庆 400052; 4. 中山大学 电子与信息工程学院，广州 510006)

摘 要：针对社交网络中社交关系的有向性与多样性，提出了一种基于图聚类与蚁群算法的社交网络聚类算法。首先，在网络覆盖率的约束下为社交网络建立有向、非全连接的二维图模型；然后，采用 K-medoids 算法搜索用户分组的中心用户，采用人工蚁群算法在 2D 图中搜索各个用户与中心用户的相似性，将满足相似性阈值的用户分为同一个用户组。设计了低活跃用户的预测机制解决网络的稀疏性问题与冷启动问题。此外，通过网络覆盖率的约束条件权衡聚类准确率与覆盖率两个指标。仿真实验结果表明，该算法实现了较好的社交网络聚类性能，并且有效地缓解了稀疏性问题与冷启动问题。

关键词：社交网络；数据挖掘；聚类处理；人工蚁群优化；图聚类；信任信息

中图分类号：TP393 doi: 10.19734/j.issn.1001-3695.2018.12.0881

Clustering algorithm of social networks based on graph clustering and ant colony optimization algorithm

Ye Xiaoying[1], Wan Mei[2], Tang Rong[3], Xie Yun[1], Chen Guihong[4], Li Qiang[1]

(1. *Dept. of Computer Science & Technology*, *Neusoft Institute of Guangdong*, *Foshan Guangdong* 528225, *China*; 2. *Dept. of Computer Science & Engineering*, *Guangzhou College of Technology & Business*, *Guangzhou* 510850, *China*; 3. *Jiulongpo District Mental Health Center*, *Chongqing* 400052, *China*; 4. *School of Electronics & Information Technology*, *Sun Yat-sen University*, *Guangzhou* 510006, *China*)

Abstract: Aiming at the properties of direction and diversity of social relationships in the social networks, this paper proposed a clustering algorithm of social networks based on graph clustering and ant colony optimization algorithm. Firstly, it constructed a directed and non fully connected complete graph for the social networks under constraint condition of network coverage; then, it adopted K-medoids algorithm to search the center users of all user groups, and it adopted ant colony optimization to search the similarities of each user and center users in the graph, it grouped the users satisfied the threshold condition into the same group. This paper also designed a prediction mechanism of low active degree users to resolve the sparsity problem and cold-start problem, besides, the network coverage constraint condition was set to balance the indexes of accuracy and coverage. Simulation experimental results indicate that the proposed algorithm realizes a good clustering performance of social networks, and it reduces the problems of sparsity and cold-start effectively.

Key words: social networks; data mining; clustering process; ant colony optimization; graph clustering; trust information

0 引言

随着微博、微信、豆瓣电影以及网易云音乐等各种应用的普及，导致不同领域的社交网络飞速地发展。目前的社交网络中存在多种社交关系，如好友关系、关注关系、具有相同喜好等[1]。社交网络的节点与连接也存在多样化的属性，传统的网络聚类方法主要考虑链接的稠密度，并未考虑社交网络的多样性。此外，社交网络中低活跃度用户的存在也为社交网络聚类效果带来了不利的影响[2]。

除了基于链接稠密度的社交网络聚类算法[3,4]，目前也出现了考虑节点多样性、强弱社交关系以及各种隐藏信息的聚类算法。文献[5]主要考虑用户的兴趣相似度，基于贝叶斯概率模型计算用户兴趣的相似度。在目前多样化的社交网络中，兴趣已经成为了一种弱关联信息，此外还应当考虑信任传播、评论信息、评分信息等。文献[6]提出了一种基于结构相似度的有向网络聚类算法，针对社交网络的有向交互性，该算法考虑了节点的到达邻居，并且采用有向边定义直接结构可达性。文献[7]采用粒子群优化算法对社交网络进行寻优处理，将网络结构作为粒子群的目标函数，通过贪婪策略引导粒子群的演化过程。文献[6,7]均将社交网络结构作为聚类的依据，但是在网络构建过程中仅考虑了直接的社交关系。

当前的社交网络中存在多样化的关联性，除了强关系，还应当考虑各种弱关系，包括关注关系、信任传播[8]、评论信息、评分信息等。此外，社交网络中存在活跃用户与低活跃度用户，而低活跃度用户会导致稀疏性问题，进而影响聚类的准确率与覆盖率[9]。为了解决上述问题，提出一种基于

收稿日期：2018-12-12；**修回日期：**2019-01-25 **基金项目：**广东省科技计划协同创新与平台环境建设基金资助项目（2017A040406001）；广东省教育厅与思科（中国）创新科技有限公司产学合作协同育人项目（粤教高函［2017］153 号）

作者简介：叶小莺（1981-），女，湖南长沙人，硕士，讲师，主要研究方向为软件工程、大数据(xueyuan___yuan@yeah.net)；万梅（1981-），女，湖南长沙人，硕士，讲师，主要研究方向为大数据、交互设计；唐蓉（1982-），女，四川罗江人，硕士，工程师，主要研究方向为电子政务、智慧城市、大数据分析；谢云（1978-），女，湖南娄底人，本科，讲师，主要研究方向为移动应用开发；陈桂宏（1983-），女，四川宜宾人，硕士，讲师，主要研究方向为无线通信安全；李强（1976-），女，湖南双峰人，硕士，副教授，主要研究方向为软件技术、数据库技术、大数据技术.

图聚类与蚁群算法的社交网络聚类算法(graph clustering and ant colony optimization social networks clustering algorithm, GC-ACO)。在覆盖率的约束下建立二维图，从而保证覆盖率与聚类准确率两者之间的平衡。在图的构建过程中，考虑了直接信任关系、信任传播、评论信息等多样化信息。结合皮尔森相似性与多样化的社交关系，以期解决稀疏性问题，设计了低活跃度用户的预测机制，以期解决冷启动问题。在聚类阶段，通过 ACO 算法搜索与中心用户相似性最高的用户，提高聚类的准确率。

1 二维图模型

在覆盖率约束下建立二维图能够有效地缓解社交网络的稀疏性问题。相似性度量的效果高度依赖用户的评论信息，因此提高相似性度量的可靠性，能够提高聚类的可靠性与精度。

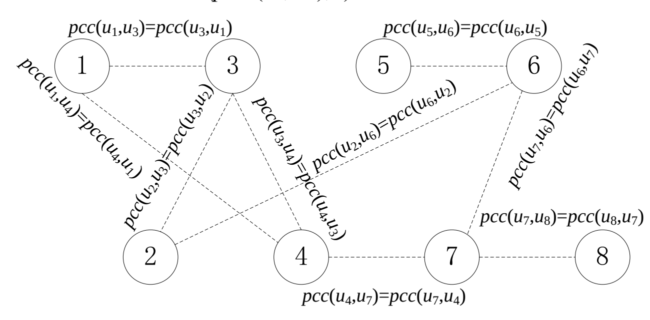
信任感知的社交网络通过预测低活跃用户的信息以提高聚类的准确率。基本思想是假设用户容易被其信任度高的用户所影响，但是该机制容易导致覆盖率降低。许多研究人员发现，用户不仅受直接信任用户影响，而且也会受间接用户的影响，但其影响力随着两个用户的距离增加而减少，该理论也称为信任传播。

设计了基于信任与相似性的图模型，该模型的建立算法如算法 1 所示。图的节点表示用户；边表示用户之间的连接；连接为双权重连接，表示为元组$(W1,W2)=(pcc(u, v), T(u, v))$，其中 pcc 表示相似性度量，T 表示信任传播。算法的输入为直接信任信息、间接信任信息、皮尔逊相关系数(Pearson correlation coefficient, PCC)以及信任传播最大距离(MP)，输出为社交网络的图。采用邻接矩阵表示社交图。根据用户之间的最短路径计算信任传播，*setdiff*()函数取消已存在的新连接。第 6 行的系数 $1/i$ 表示两个用户距离越长，其信任值越低。

算法 1：社交网络的图建立算法

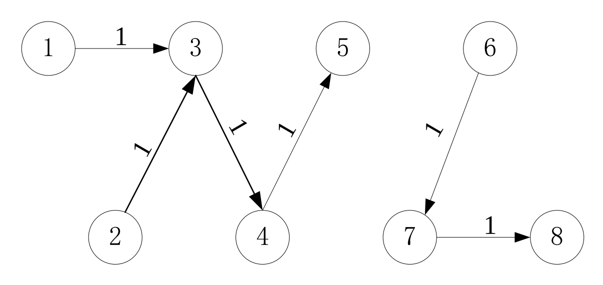
```
输入：PCC 图，信任图，MP      /*MP 为信任传播最大距离*/。
输出：W_图。
1.uers=用户数量；
2.tmp=I_users×users；          /*初始化临时用户矩阵*/
3.mt=0_users×users；       /*初始化用户矩阵*/
4.foreach i=1 to MP do {
5.    tmp = tmp * T;
6.    mt = mt +(1/i)setdiff(mt, tmp)；   /*计算信任用户之间的差异*/
7.}
8.foreach (u_i, u_j) do {      //遍历每对用户
9.    if PCC(u_i, u_j)与MT(u_i, u_j)两者均存在 { /*同时存在 PCC 图与 MT 值*/
10.        (W_s, W_mt) = (PCC(u_i, u_j), 0);
11.    } else if 存在 PCC(u_i, u_j) {
12.        (W_s, W_mt)=(PCC(u_i, u_j), 0);
13.    } else if 存在 MT(u_i, u_j) {
14.        (W_s, W_mt)=(0, MT(u_i, u_j));
15.    }
16.    W_图(u_i, u_j)=(W_s, W_mt);
17.}
```

图 1 所示是包含八个用户的社交网络实例。其中图 1(a)是 PCC 相似性图；(b)是社交网络的直接信任关系图；(c)是社交网络的信任传播图；(d)是最终建立的图模型。从图 1 可看出，社交网络图可能不是一个全连接图，根据六度空间理论[10]，两个连接用户之间最长距离为六跳。将图 1(a)、(b)、(c)三者组成图 1(d)的图，该集成程序可能为孤立的分区建立连接，该程序有助于提高覆盖率。图中 u 表示用户，图 1(b)、(c)中边的值为两个用户之间的 pcc 值。图中 u_1 信任 u_3，u_3 不信任 u_1，$u_1 \rightarrow u_3$ 的权重为 $(pcc(u_1,u_3),T(u_1,u_3))=(pcc(u_1,u_3),1)$，而 $u_3 \rightarrow u_1$ 的权重为$(pcc(u_3, u_1),0)$，因此图为有向图。



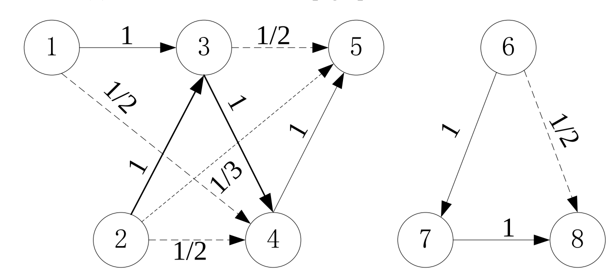
(a) PCC 相似性图

(a) PCC similarity graph



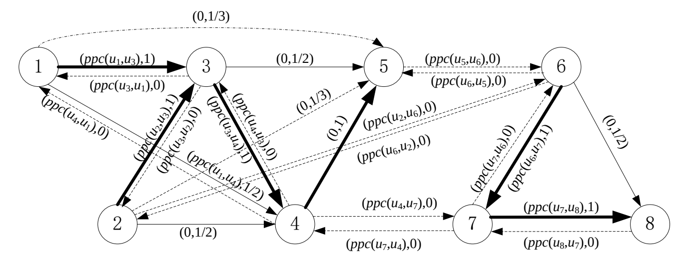
(b) 社交网络的直接信任关系图

(b) Direct trust relationship graph of social networks



(c) 社交网络的信任传播图

(c) Trust propagation graph of social networks



(d) 最终建立的二维图模型

(d) Finally constructed two dimensional graph model

图 1 包含八个用户的社交网络实例

Fig. 1 Social networks example including eight users

2 蚁群算法的背景知识

人工蚁群算法(ant colony optimization, ACO)[11]是一种多 agent 系统，能够分布式地求解问题，并且具有较强的全局搜索能力与局部开发能力。蚁群算法首先将问题建模为一个加权图，然后搜索图中的最优路径。人工蚂蚁通过游走产生可

行解，蚂蚁间互相交换信息，并且在边或者节点释放信息素。ACO 中蚂蚁 k 释放信息素的过程可定义为

$$\tau_{ab}=\begin{cases}(1-\rho)\tau_{ab}+\dfrac{Q}{L_k}, & if\ \text{蚂蚁}k\text{使用曲线}ab\text{在其路径中}\\(1-\rho)\tau_{ab}, & \text{其他情况}\end{cases} \tag{1}$$

其中：τ_{ab} 为蚂蚁在边 e_{ab} 上释放的信息素；ρ 为信息素的挥发系数；L_k 为蚂蚁 k 搜索解的成本；Q 为一个常量。为了防止 ACO 陷入局部最优，信息素应当随着时间挥发，从而提高探索新解的可能性。人工蚂蚁在游走过程中选择信息素较高的路径，蚂蚁 k 从状态 a 变为状态 b 的概率为

$$p_{ab}^{k}=\frac{\tau_{ab}^{\alpha}\eta_{ab}^{\beta}}{\sum_{z\in allowed_{\alpha}}\tau_{az}^{\alpha}\eta_{az}^{\beta}} \tag{2}$$

其中：α 为控制 τ_{ab} 影响力的参数；η_{ab} 为状态 a 变为状态 b 的可能性；β 为控制 η_{ab} 影响力的参数。

3　本文的算法设计

3.1　产生合适的用户分组数量

社交网络的覆盖率与分组数量没有相关性，因此，可将搜索出的第一个分组数量作为聚类算法的分组数。

3.2　确定用户分组的中心用户

采用 K-medoids 算法[12]搜索用户分组的中心用户。K-medoids 算法的目标函数(F)定义为

$$F=\min\sum_{c_E\in C}\sum_{m,n\in C_E}dist(m,n) \tag{3}$$

其中：C 为类的集合；$dist(m, n)$表示二维图中用户 m 与 n 的距离。因为图中每条边为双权重，所以用户间的距离计算为

$$dist^2(u,v)=d_S^2(u,v)+d_T^2(u,v) \tag{4}$$

其中：u 与 v 为两个目标用户；d_S 为相似性距离，计算方法为

$$d_S(u,v)=1-W_S^{2D\text{-}Graph}(u,v) \tag{5}$$

d_T 为信任距离，计算方法为

$$d_T(u,v)=1-W_{MT}^{2D_Graph}(u,v) \tag{6}$$

3.3　基于 ACO 的社交网络聚类

3.2 节选出了用户分组的中心用户，然后寻找与中心用户相似性高的用户组。该过程主要包括排列处理、加权处理、预测处理三个步骤。

3.3.1　初始化排列处理

该步骤的目标是基于信任信息与评论信息计算各个用户与目标用户(中心用户)之间的相似性值，提取出 top-n 的相似用户。如果用户之间存在直接信任关系，如好友关系、关注关系等，那么直接计算信任值；如果用户之间不存在直接信任关系，那么根据提取隐藏的信任关系，如评论信息、评分信息等。如果用户 u 与目标用户 a 之间不存在直接的信任关系，使用 PCC 根据评论信息或者评分信息计算 u 与 a 的信任值，网络的节点表示用户，边的权重表示用户之间的相似性。基于信任的用户相似性计算[13]为

$$W_{a,u}=\begin{cases}\dfrac{2\times sim(a,u)\times T_{a,u}}{sim(a,u)+T_{a,u}}, & sim(a,u)+T_{a,u}\neq 0\\T_{a,u}, & sim(a,u)=0, T_{a,u}\neq 0\\sim(a,u), & sim(a,u)\neq 0, T_{a,u}=0\end{cases} \tag{7}$$

其中：$T_{a,u}$ 为目标用户 a 与用户 u 之间的信任值，计算式为

$$T_{a,u}=\frac{d_{max}-d_{a,u}+1}{d_{max}} \tag{8}$$

其中：$d_{a,u}$ 表示 a 与 u 之间的信任传播距离；d_{max} 为最大的信任传播距离，d_{max} 设为图中的平均路径长度。

$$d_{max}=\frac{\ln(n)}{\ln(k)} \tag{9}$$

其中：n 为网络中的用户数量；k 为网络的平均度。假设 $sim(a, u)$表示 a 与 u 的相似性，基于 PCC 的相似性计算为

$$sim(a,u)=\frac{\sum_{i\in A_{a,u}}\left(r_i(a)-\bar{r}(a)\right)\left(r_i(u)-\bar{r}(u)\right)}{\sqrt{\sum_{i\in A_{a,u}}\left(r_i(a)-\bar{r}(a)\right)^2}\sqrt{\sum_{i\in A_{a,u}}\left(r_i(u)-\bar{r}(u)\right)^2}} \tag{10}$$

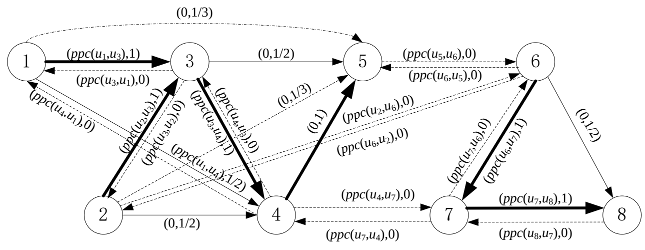
其中：$r_i(u)$为用户 u 对于项目 i 的评分值；$\bar{r}(u)$ 为用户 u 的平均评分值；$A_{a,u}$ 为用户 a 与 u 评分的项目集合。最终，将相似性高于阈值 θ 的分为一个用户组。

3.3.2　二维图模型的加权处理

采用 ACO 处理 top-n 用户，分析他们的重要性。首先，建立用户的二维图；然后，蚁群在图中游走以调节各个用户与目标用户的相似性。

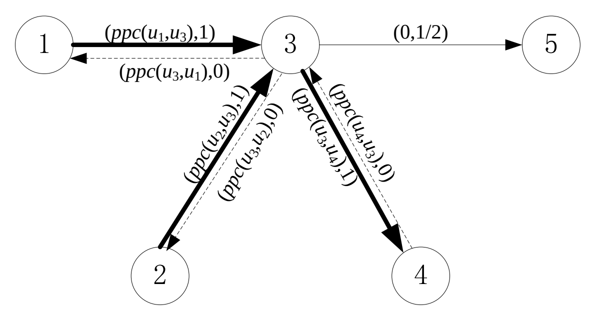
1)　建立用户二维图

首先，选择与目标用户 top-n 相似的用户；然后，为社交网络建立第 1 章的二维图，其中节点表示用户，边与权重表示用户之间的相似性(式(7)计算)，权重的取值为[0,1]。图 2(a)所示是一个社交网络的二维图例子；(b)所示是目标用户 3 的子图。启发式信息与期望信息是 ACO 算法的两个主要元素，启发式信息定义为反权重值。



(a) 社交网络的二维图例子

(a) Two dimensional graph example of social networks



(b) 目标用户 3 的子图

(b) Sub-graph of target user 3

图 2　提取中心用户子网的实例

Fig. 2　Example of sub-network of center user abstraction

2)　蚁群游走策略

初始化阶段，将蚁群随机置于图中，然后蚁群在游走过程中更新信息素。蚂蚁根据用户与目标用户的相似性释放合适的信息素量，蚂蚁基于一个路由表在图中游走。蚂蚁 k 从节点 i 移至节点 j 的概率定义为

$$P_k(i,j)=\begin{cases}\dfrac{[\tau_j]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{m\in N_i^k\wedge unvisited(m)}[\tau_m]^{\alpha}[\eta_{im}]^{\beta}}, & if\ j\in N_i^k\wedge\eta_{ij}\leq\Omega\\0, & if\ j\notin N_i^k\end{cases} \tag{11}$$

其中：N_i^k 为节点 i 的邻居集；τ 为信息素量；η 为启发值；α 与 β 分别为控制 τ 与 η 权重的参数；$\eta_{ij}=1/sim(u_i, u_j)$；$m$ 为尚未访问的用户。式(11)的概率函数能够防止算法陷入局部最

优，在社交网络中该函数能够选择与目标用户兴趣相似、冗余度低的用户集。

3) 信息素更新方法

ACO 中信息素反映了蚂蚁求解一个问题的经历，信息素的更新反映了蚂蚁的解质量。文中节点的信息素表示用户与目标用户的相关性，用户 u_i 信息素的更新方法为

$$\tau_{u_i} = \tau_{u_i} + \sum_k \Delta\tau_{u_i}^k \tag{12}$$

其中：τ_{u_i} 为用户 u_i 的信息素；$\Delta\tau_{u_i}^k$ 表示蚂蚁 k 在用户 i 释放的信息素量。$\Delta\tau_{u_i}^k$ 反映了解的质量，计算方法为

$$\Delta\tau_{u_i}^k = \begin{cases} \dfrac{Q}{cost(U^k)}, & u_i \in U^k \\ 0, & u_i \notin U^k \end{cases} \tag{13}$$

其中：Q 为常量；U_k 为蚂蚁 k 经历的用户集；$cost(U^k)$为蚂蚁 k 发现解的质量。每次迭代结束，更新所有节点的信息素：

$$\tau_{u_i} = \tau_{u_i}(1-\rho) \tag{14}$$

其中：ρ 为信息素的挥发速率。采用平均误差指标计算每个解的质量，每个解由一个用户集及其权重组成。

3.3.3 低活跃用户的预测处理

对于缺少评论信息的用户，根据与其最相似的用户评论预测其对目标用户的评论。预测方法为

$$\hat{r}_{u,i} = \frac{\sum_{v\in U} w_v r_{v,i}}{\sum_{v\in U} w_v} \tag{15}$$

其中：$\hat{r}_{u,i}$ 为目标用户 u 对于项目 i 的预测评论；U 为蚂蚁 k 选择的用户集；$r_{v,i}$ 为 v 对项目 i 的真实评分；w_v 为 v 的信息素。每个解的成本计算为预测值与真实值之间的误差。

$$fitness(u) = \frac{\sum_{i=1}^{I_u} \left|\hat{r}_{u,i} - r_{u,i}\right|}{|I_u|} \tag{16}$$

其中：I_u 为预测的项目数量。

该处理的目标是根据活跃用户的信息预测低活跃用户的信息，该处理有助于缓解社交网络中普遍存在的冷启动问题、稀疏性问题等。

3.4 算法总体设计

初始化阶段，每个节点的信息素设为常量 c，蚁群随机置于图中各节点的位置。每个蚂蚁基于式(11)在图中游走，蚂蚁可能选择不同数量的用户集。蚂蚁根据(12)式更新各个用户的信息素。考虑信息素的挥发，在每次迭代的结束阶段根据式(14)进行信息素的全局更新。重复上述步骤，直至达到结束条件。将用户按重要性降序排列，选择 top-k 的用户作为最终的子集。

算法 2 所示是 GC-ACO 算法的伪代码。算法的输入变量 R、T、m、N_{ant}、NI 分别表示评论信息(评分)、信任信息、用户数量、目标用户、蚁群规模、迭代次数。算法步骤如下：a)计算目标用户与其他用户的相似性，选择相似性高于 θ 的用户输入 ACO 算法处理；b)采用 ACO 为用户分配权重，ACO 的每次迭代中，蚂蚁在图中游走，选择目标用户的一个相似用户集，步骤 b)的输出是一个包含信息素值的用户集；c)通过预测程序提高低活跃用户的聚类效果与覆盖率。

算法 2 基于二维图与 ACO 的社交网络聚类算法

```
输入：R，T，m，Nant，NI。
输出：与目标用户相似的用户集。
/*基于评论信息与信任信息初步筛选用户*/
1.计算目标用户 a 与其他用户的相似性；//(1)式
2.SU=选择相似性高于阈值 θ 的用户集；
/*采用 ACO 计算用户的权重*/
3.建立用户的二维图；
4.初始化图中节点的信息素；
5.foreach i = 1 to NI {
6.    随机分布蚁群；
7.    foreach j = 1 to Nant {
8.        Uk = []；
9.        while (如果当前用户的启发值低于阈值) {
10.           选择下一个未访问的用户 u；// 式(11)
11.           将 u 加入向量 Uk 中；
12.       }
13.       计算适应度；//式(16)
14.       更新信息素；//式(12)
15.   }
16.   更新全局信息素；//式(14)
17.}
/*低活跃用户的预测*/
18.基于信息素将用户降序排列；
19.选择 top-k 用户；
20.预测 a 的未知评价；
```

3.5 GC-ACO 算法的复杂度分析

算法 2 的步骤 a)，因为每对用户之间的相似性依赖用户的总数量，所以计算复杂度为 $O(n^2)$，第二行选择相似性高于 θ 的用户，设为 $l=|SU|$，建立包含 l 个用户的二维图。步骤 b)，该步骤的迭代次数为 NI，其计算复杂度为 $O(NI \cdot N_{ant} \cdot l^2)$，如果采用分布式处理，那么该步骤的复杂度可降为 $O(NI \cdot l^2)$。步骤 c)，该步骤的计算复杂度为 $O(l \log l)$。最终，本算法的总体计算复杂度为 $O(n^2+NI \cdot l^2+l \log l)$。

4 实验与结果分析

推荐系统是社交网络聚类技术的一个重要应用场景，采用文献[5,14]的实验方案，将聚类技术与协同过滤推荐系统结合，通过推荐系统的效果评估社交网络推荐技术的效果。采用三个数据集测试 GC-ACO 算法的聚类性能。实验环境为 PC 机，PC 机的配置为 8 GB 内存，i7 8700 处理器。采用五折交叉检验的实验方案，将每个数据集分为五个子集，每次迭代中随机选择四个子集作为训练集，另外一个子集作为测试集。

4.1 性能评价指标

采用三个经典的推荐系统性能指标，即均方误差(MAE)、根均方误差(RMSE)、覆盖率(RC)。MAE 用于评估预测的准确率。MAE 计算预测评分值与真实评分值之间的差异。

$$MAE=\frac{1}{Z}\sum_{(u,j)}\left|\hat{r}_{uj}-r_{uj}\right| \tag{17}$$

其中：Z、$\hat{r}_{uj}$ 与 r_{uj} 分别为用户 u 对于项目 j 的评分数量、估计评分数量以及真实评分数量。RMSE 也是评估推荐系统性能的一个指标，该指标度量了预测评分与真实评分的绝对误差。

$$RMSE=\sqrt{\frac{1}{Z}\sum_{(u,j)}\left(\hat{r}_{uj}-r_{uj}\right)^2} \tag{18}$$

RC 从另一个角度评估推荐系统的性能，评估了推荐系统对长尾商品的挖掘能力。RC 的计算方法为

$$RC=\frac{\text{预测的评分数量}}{\text{所有的评分数量}} \tag{19}$$

4.2　实验数据集

采用三个数据集作为 benchmarks 数据集，分别为 FilmTrust、Epinions、Ciao 数据集。FilmTrust 是一个电影推荐网站的真实数据集，该网站的用户对电影进行评论与评分，用户之间也可添加好友并分享观点。FilmTrust 数据集的评分为实数，范围为 0.5~4。Epinions 数据集包括多种社交关系，包括对项目的评论与评分以及用户之间的信任关系，评分为整数，范围为 1~5；信任关系为两个值："1"表示信任，"0"表示不信任。Ciao 数据集的评分为整数，范围为 1~5。三个 benchmark 数据集的相关信息如表 1 所示。

表 1　benchmark 数据集的相关信息

Table 1　Related information of benchmark datasets

特征	FilmTrust	Epinions	Ciao
用户	1508	40163	30444
项目	2071	139738	72665
评分	35497	664824	1625480
信任者	609	33960	6792
受信任者	732	49288	7297
信任量	1853	487183	111781

为了测试本算法对稀疏性问题、冷启动问题的效果，按照两种条件进一步划分数据集，划分条件为：a)冷启动用户，提取评分数量少于 5 的用户集；b)稀疏性项目，提取评分数量少于 5 的项目；c)全部用户集。表 2 所示是划分子数据集的相关信息。

表 2　划分子数据集的相关信息

Table 2　Related information of datasets division

划分条件	数据集	实例数量	评分数量
冷启动	FilmTrust	281	608
	Epinions	16910	33632
	Ciao	12006	20985
稀疏性	FilmTrust	1653	3162
	Epinions	116152	175906
	Ciao	9423	24722

4.3　参数设置

通过多组预处理实验选择出最优的参数配置：最大循环次数设为 70，初始化信息素与信息素挥发系数分别设为 0.02 与 0.2。参数 α、β、Ω 分别设为 α=0.6, β=0.4, Ω=1.66，用户的邻居数量设为 2~30。蚁群的蚂蚁数量等于各个数据集的用户数量。

4.4　实验结果

选择近期两个基于社交网络的推荐系统与一个基于智能优化的推荐系统作为对比算法，分别为：a)基于信任与用户评分的推荐系统 TrustSVD[15]；b)基于信任与矩阵分解的推荐系统 TrustMF[16]；c)基于遗传算法的推荐系统 Yilmaz[17]。

4.5　不同数据集的推荐性能

将 TrustSVD、TrustMF、Yilmaz 与 GC-ACO 四种算法对冷启动数据集、稀疏数据集以及完整数据集进行了推荐实验，统计每组实验的 MAE 与 RMSE 指标的结果，表 3~5 分别是 FilmTrust、Epinions 与 Ciao 数据集的实验结果。GC-ACO 算法对于 FilmTrust、Epinions 两个数据集的准确率较好，优于其他三个推荐系统。对于 Ciao 数据集也取得了较好的结果，但其对完整数据集的推荐准确率略低于 TrustMF 系统，对稀疏数据集的推荐准确率略低于 TrustSVD 系统。总体而言，GC-ACO 取得了较好的推荐效果，对于冷启动问题与稀疏性问题均实现了加好的缓解效果。

表 3　推荐系统对于 FilmTrust 数据集的 MAE 与 RMSE 结果

Table 3　MAE and RMSE results of recommender systems applied to filmtrust dataset

数据集	指标	TrustSVD	TrustMF	Yilmaz	GC-ACO
冷启动	MAE	0.650	0.619	0.722	**0.586**
	RMSE	0.845	0.882	0.931	**0.758**
稀疏集	MAE	0.829	0.907	0.841	**0.793**
	RMSE	1.059	1.249	1.084	**1.003**
完整集	MAE	0.607	0.721	0.685	**0.496**
	RMSE	0.787	0.919	0.912	**0.721**

表 4　推荐系统对于 Epinions 数据集的 MAE 与 RMSE 结果

Table 4　MAE and RMSE results of recommender systems applied to Epinions dataset

数据集	指标	TrustSVD	TrustMF	Yilmaz	GC-ACO
冷启动	MAE	0.861	0.934	0.871	**0.795**
	RMSE	1.117	1.373	1.124	**1.026**
稀疏集	MAE	0.829	0.856	0.824	**0.801**
	RMSE	1.096	1.19	1.082	**1.033**
完整集	MAE	0.834	0.877	0.852	**0.769**
	RMSE	1.094	1.184	1.101	**1.021**

表 5　推荐系统对于 Ciao 数据集的 MAE 与 RMSE 结果

Table 5　MAE and RMSE results of recommender systems applied to Ciao dataset

数据集	指标	TrustSVD	TrustMF	Yilmaz	GC-ACO
冷启动	MAE	0.725	1.073	0.747	**0.688**
	RMSE	0.939	1.311	0.932	**0.903**
稀疏集	MAE	**0.503**	1.209	0.532	0.516
	RMSE	0.659	1.493	0.675	**0.632**
完整集	MAE	0.723	0.505	**0.491**	0.503
	RMSE	0.955	**0.493**	0.670	0.659

覆盖率指标是推荐系统与社交网络的重要指标，统计了四个推荐系统的覆盖率结果，结果如图 3 所示。从图中可看出，四种算法均实现了较高的覆盖率，TrustMF、Yilmaz 与 GC-ACO 三个推荐系统的覆盖率均高于 0.9，而本算法的覆盖率则略高于 TrustMF 与 Yilmaz 算法。

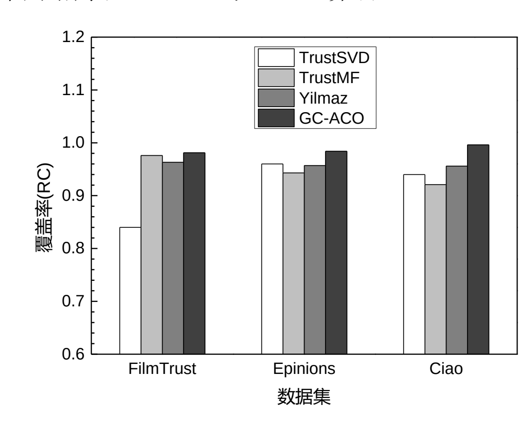


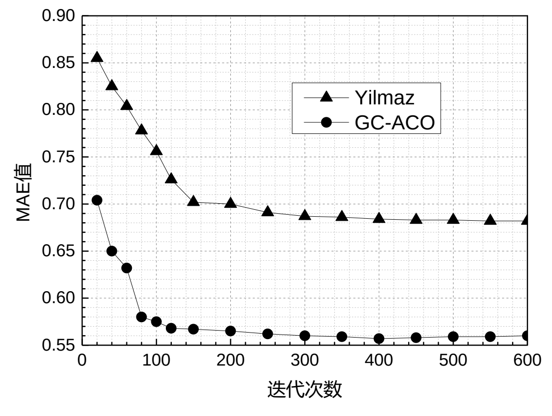
图 3　不同推荐系统的覆盖率结果

Fig. 3　Coverage rate results of different recommender systems

4.6　收敛性分析

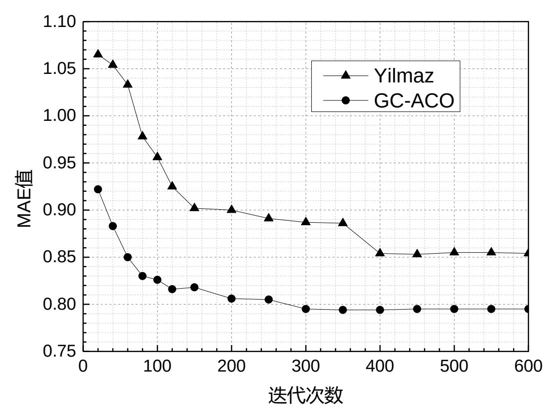
TrustSVD 与 TrustMF 并非基于迭代的算法，Yilmaz 系统基于遗传算法算法实现，将本算法与 Yilmaz 算法比较，观察本算法的收敛性。本算法是一种基于迭代的算法，收敛性是迭代性算法的关键性能。Yilmaz 是一种基于遗传算法的推荐系统，将本算法与 Yilmaz 算法进行比较，两种算法对于

FilmTrust、Epinions 与 Ciao 数据集的收敛曲线分别如图4(a)~(c)所示。从图中可看出，本算法的收敛速度与准确率结果均优于 Yilmaz 算法。本算法的全局搜索能力较强，实现了较好的准确率，局部开发能力较强，实现了较快的收敛速度。式(11)的概率函数防止蚁群算法陷入局部最优，从而实现了较强的全局搜索能力；式(11)未使用贪婪机制，使得低概率用户依然具有被选择的可能性。另外，第一步对用户进行了初步筛选，使得本算法保持了较高的开发能力，并且缩小了解空间。本算法采用了丰富的直接信任关系与间接信任关系建立图中的权重，该机制使蚁群在迭代初期即可快速、高效地在图中游走，因此本算法实现了较好的开发能力与收敛速度。



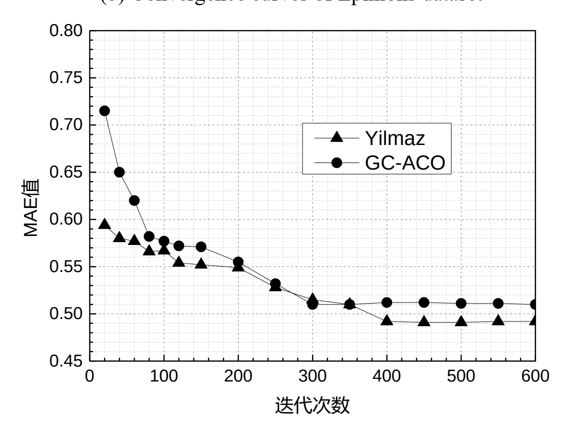
(a) FilmTrust 数据集的收敛曲线

(a) Convergence curves of FilmTrust dataset



(b) Epinions 数据集的收敛曲线

(b) Convergence curves of Epinions dataset



(c) Ciao 数据集的收敛曲线

(c) Convergence curves of Ciao dataset

图 4　两个基于迭代推荐系统的收敛性曲线

Fig. 4　Coverage curves of two iteration based recommender systems

5　结束语

针对社交网络中社交关系的有向性与多样性，本文提出了一种基于图聚类与蚁群算法的社交网络聚类算法。在覆盖率的约束下建立二维图，从而保证覆盖率与聚类准确率两者之间的平衡。在图的构建过程中，考虑了直接信任关系、信任传播、评论信息等多样化信息。本算法取得了较好的推荐效果，对于冷启动问题与稀疏性问题均实现了较好的缓解效果。本算法采用了丰富的直接信任关系与间接信任关系建立图中的权重，该机制使蚁群在迭代初期即可快速、高效地在图中游走，因此本算法实现了较好的开发能力与收敛速度。未来将考虑引入更多的隐藏社交信息与外部信息以增强社交网络的判断依据，如用户档案、评论上下文以及行为轨迹等信息。

参考文献：

[1] 胡长军，许文文，胡颖，等. 在线社交网络信息传播研究综述 [J]. 电子与信息学报, 2017, 39 (4): 794-804. (Hu Changjun, Xu Wenwen, Hu Ying, *et al.* Review of information diffusion in online social networks [J]. Journal of Electronics & Information Technology, 2017, 39 (4): 794-804.)

[2] 赵姝，刘晓曼，段震，等. 社交关系挖掘研究综述 [J]. 计算机学报, 2017, 40 (3): 535-555. (Zhao Shu, Liu Xiaoman, Duan Zhen, *et al.* A survey on social ties mining [J]. Chinese Journal of Computers, 2017, 40 (3): 535-555.)

[3] Stovall T R, Kockara S, Avci R. GPUSCAN: GPU-based parallel structural clustering algorithm for networks [J]. IEEE Trans on Parallel & Distributed Systems, 2015, 26 (12): 3381-3393.

[4] Zhao Weizhong, Chen Gang, Xu Xiaowei. AnySCAN: an efficient anytime framework with active learning for large-scale network clustering [C]// Proc of IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2017: 665-674.

[5] 汤颖，钟南江，孙康高，等. 基于兴趣的社交网络用户聚类及可视化 [J]. 计算机科学, 2017, 44 (b11): 385-390. (Tang Ying, Zhong Nanjiang, Sun Kanggao, *et al.* Clustering and visualization of social network based on user interests [J]. Computer Science, 2017, 44 (b11): 385-390.)

[6] 陈季梦，陈佳俊，刘杰，等. 基于结构相似度的大规模社交网络聚类算法 [J]. 电子与信息学报, 2015, 37 (2): 449-454. (Chen Jimeng, Chen Jiajun, Liu Jie, *et al.* Clustering algorithms for large-scale social networks based on structural similarity [J]. Journal of Electronics & Information Technology, 2015, 37 (2): 449-454.)

[7] Cai Qing, Gong Maoguo, Ma Lijia, *et al.* Greedy discrete particle swarm optimization for large-scale social network clustering [J]. Information Sciences An International Journal, 2015, 316 (9): 503-516.

[8] Wu Jian, Chiclana Francisco, Fujita Hamido, *et al.* A visual interaction consensus model for social network group decision making with trust propagation [J]. Knowledge-Based Systems, 2017, 122 (C): 39-50.

[9] 孟祥武，刘树栋，张玉洁，等. 社会化推荐系统研究 [J]. 软件学报, 2015, 26 (6): 1356-1372. (Meng Xiangwu, Liu Shudong, Zhang Yujie, *et al.* Research on social recommender systems [J]. Journal of Software, 2015, 26 (6): 1356-1372.)

[10] 刘宏杰，陆浩，张楠，等. 基于微博的六度空间理论研究 [J]. 计算机应用研究, 2012, 29 (8): 2826-2829. (Liu Hongjie, Lu Hao, Zhang Nan, *et al.* Theory research based on microblog of six degrees space [J]. Application Research of Computers, 2012, 29 (8): 2826-2829.)

[11] Eaton Jayne, Yang Shengxiang, Mavrovouniotis M. Ant colony optimization with immigrants schemes for the dynamic railway junction rescheduling problem with multiple delays [J]. Soft Computing, 2016, 20 (8): 2951-2966.

[12] 韩啸，刘淑芬，徐天琦. 基于遗传模拟退火算法的改进 K-medoids 算法 [J]. 吉林大学学报 ：工学版 , 2015, 45 (2): 619-623. (Han Xiao, Liu Shufen, Xu Tianqi. Improved K-medoids algorithm based on genetic simulated annealing algorithm [J]. Journal of Jilin University ：Engineering and Technology Edition , 2015, 45 (2): 619-623.)

[13] Moradi P, Ahmadian S. A reliability-based recommendation method to improve trust-aware recommender systems [J]. Expert Systems with Applications, 2015, 42 (21): 7386-7398.

[14] 陈克寒，韩盼盼，吴健. 基于用户聚类的异构社交网络推荐算法 [J]. 计算机学报, 2013, 36 (2): 349-359. (Chen Kehan, Han Panpan, Wu Jian. User clustering based social network recommendation [J]. Chinese Journal of Computers, 2013, 36 (2): 349-359.)

[15] Guo Ging, Zhang Jie, Yorke-Smith Neil. TrustSVD: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings [C]// Proc of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 123-129.

[16] Yang Bo, Lei Yu, Liu Jiming, *et al.* Social collaborative filtering by trust [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (8): 1633-1647.

[17] Ar Y, Bostanci E. A genetic algorithm solution to the collaborative filtering problem [J]. Expert Systems with Applications, 2016, 61 (1): 122-128.